

# Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides Derived from Cellular Proteins

**Jacek P. Dworzanski**

Geo-Centers, Inc.,  
Aberdeen Proving Ground, MD 21010-0068

**Samir Deshpande<sup>1</sup>; Rui Chen<sup>2</sup>; A. Peter Snyder<sup>3</sup>;  
Liang Li<sup>2</sup> and Charles H. Wick<sup>3</sup>**

<sup>1</sup>Science and Technology Corporation, Edgewood, MD 21040; <sup>2</sup>Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2; <sup>3</sup>U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010-5424

**2004 Joint Service Scientific Conference on Chemical &  
Biological Defense Research  
15-17 November 2004; Hunt Valley, Maryland**

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE <b>17 NOV 2004</b>		2. REPORT TYPE <b>N/A</b>		3. DATES COVERED <b>-</b>	
4. TITLE AND SUBTITLE <b>Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides Derived from Cellular Proteins</b>				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Science and Technology Corporation, Edgewood, MD 21040</b>				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release, distribution unlimited</b>					
13. SUPPLEMENTARY NOTES <b>See also ADM001849, 2004 Scientific Conference on Chemical and Biological Defense Research. Held in Hunt Valley, Maryland on 15-17 November 2004 . , The original document contains color images.</b>					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>UU</b>	18. NUMBER OF PAGES <b>27</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

**National Institute of Allergy and Infectious Diseases  
National Institute of Health**

**Category A, B & C Priority Bacterial Pathogens**

**Category A**

1. *Clostridium botulinum*
2. *Bacillus anthracis* (anthrax)
3. *Francisella tularensis* (tularemia)
4. *Yersinia pestis*

**Category C**

1. *Mycobacterium tuberculosis*  
(multiple drug resistant)
2. *Rickettsias* (other)

**Category B**

1. *Brucella species* (brucellosis)
2. *Burkholderia pseudomallei*
3. *Burkholderia mallei* (glanders)
4. *Campylobacter jejuni*
5. *Clostridium perfringens* (epsilon toxin)
6. *Coxiella burnetti* (Q fever)
7. *Escherichia coli* (diarrheagenic)
8. *Listeria monocytogenes*
9. *Rickettsia prowazekii* (typhus fever)
10. *Salmonella*
11. *Shigella species*
12. *Staphylococcus aureus* (enterotoxin B)
13. *Vibrios* (pathogenic)
14. *Yersinia enterocolitica*

**Genomes of all above organisms have been sequenced**

5/12/2004 versus {11/12/2004}

Number of Fully Sequenced Genomes of Eubacteria: 145 {178} (From 1995)

Number of Fully Sequenced Genomes During Last 12 Months: 55 {68}

Prokaryotic Ongoing Genome Projects: 489 {528}

Archaeal: 28 {27}

Bacterial: 461 {499}

## Example of Ongoing Genome Projects

***Bacillus anthracis*** A1055 (Group C)

TIGR/NIH

***Bacillus anthracis*** Ames Ancestor

TIGR/NIH / NSF

***Bacillus anthracis*** Ames Florida

TIGR/NSF

***Bacillus anthracis*** Australia 94 (GT55, Group A3a)

TIGR/NIH

***Bacillus anthracis*** CNEVA-9066 (GT79 Group B2)

TIGR/NIH

***Bacillus anthracis*** Kruger B (GT87 Group B1)

TIGR/NIH

***Bacillus anthracis*** Vollum (GT77 Group A4)

TIGR/NIH

***Bacillus anthracis*** Western N. America (GT3 Group A1a)

TIGR/NIH

***Bacillus anthracis*** STN

DOE/JGI

***Bacillus anthracis*** ZK

DOE/JGI

# First, some terminology...

- Taxonomy** - the science of naming and classifying organisms;
- Classification** - placement of an organism within a scheme relating different groups of organisms;
- Identification** - the determination of whether an organism should be placed within a group of organisms known to fit within some classification scheme;  
(the practical use of classification criteria)
- Phylogenetics** - focuses on evolutionary relationships between organisms or genes/proteins

## Phylogenetic Approach

The ideal means of identifying and classifying bacteria would be to compare each gene sequence in a given strain with the gene sequences for every known species.

# Taxonomy of Bacteria

## (Linnaean System)

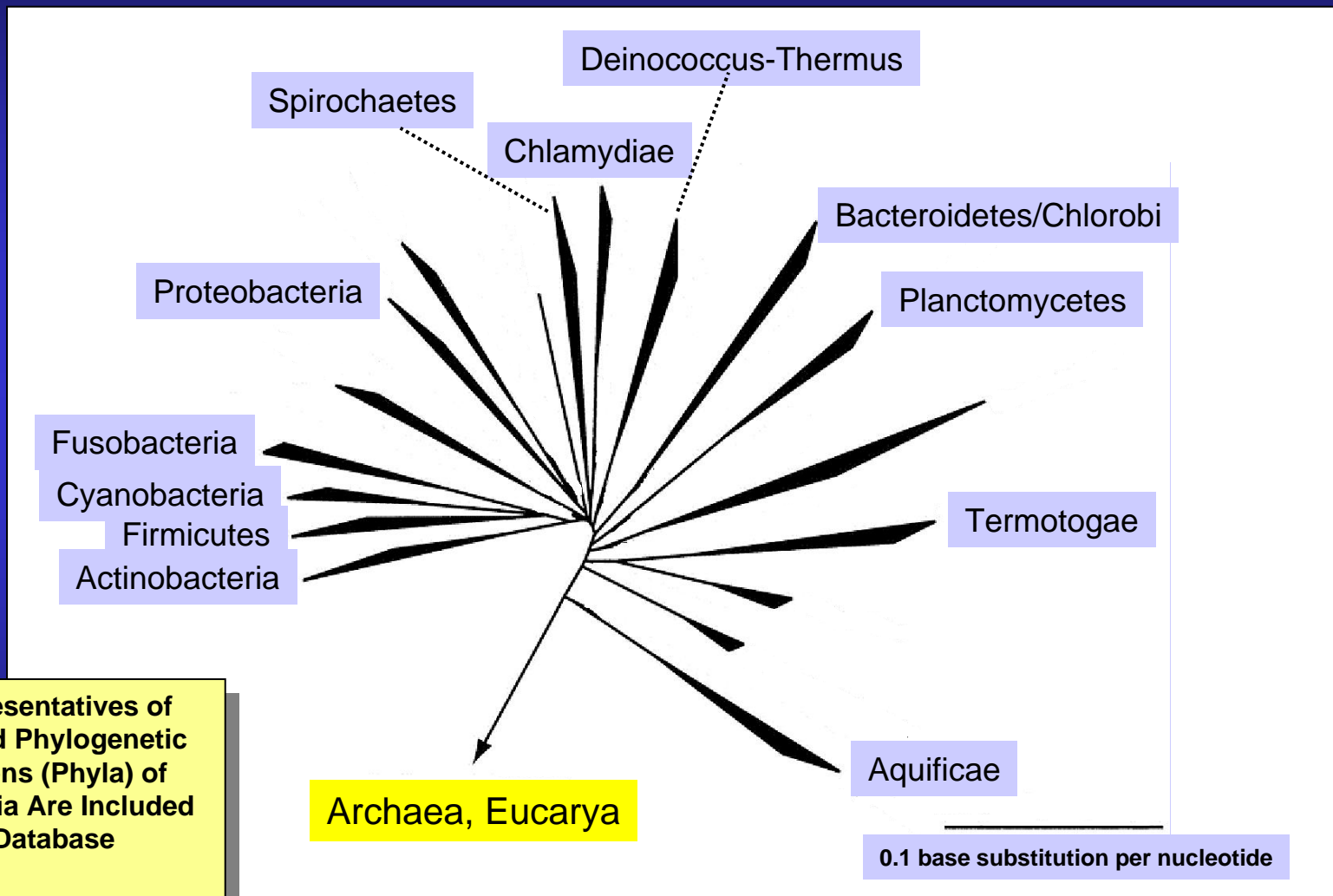
Examples: *Bacillus subtilis* & *Escherichia coli*

Kingdom  
Phylum  
Class  
Order  
Family  
Genus  
Species

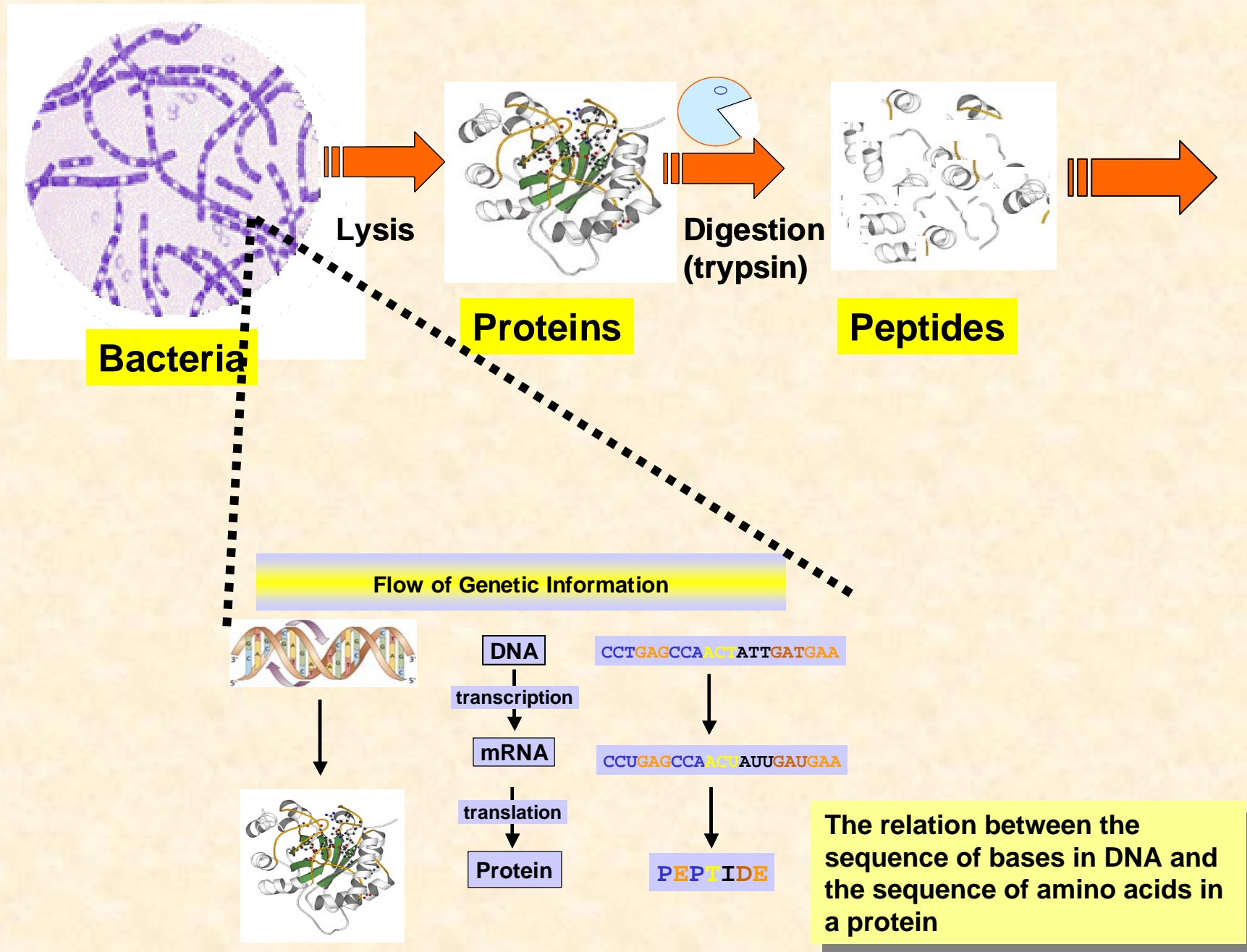
Bacteria  
Firmicutes  
Bacilli  
Bacillales  
Bacillaceae  
Bacillus  
*Bacillus subtilis*

Bacteria  
Proteobacteria  
 $\gamma$ -Proteobacteria  
Enterobacteriales  
Enterobacteriaceae  
Escherichia  
*Escherichia coli*

# Universal Phylogenetic Tree of Bacteria Based on SSU rRNA Sequences

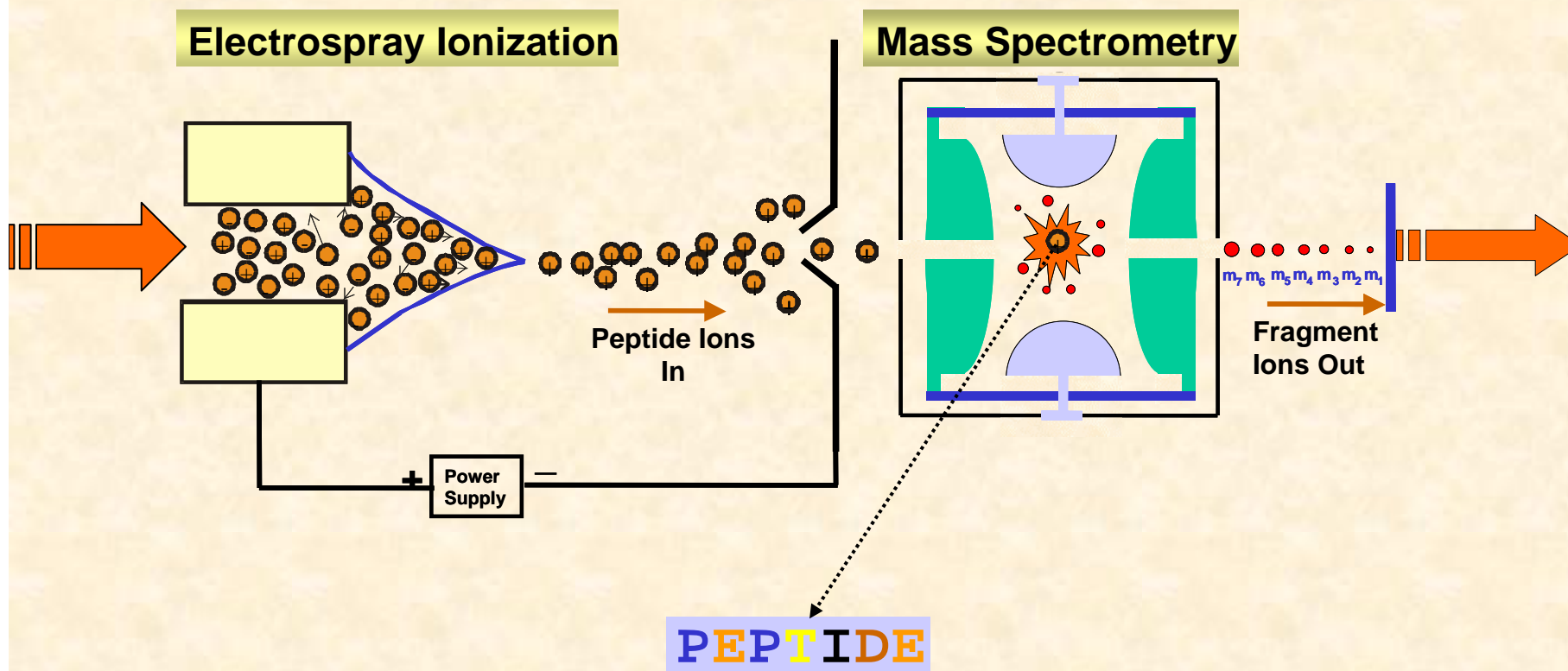


# Approach: (1) Bacterial Sample Processing



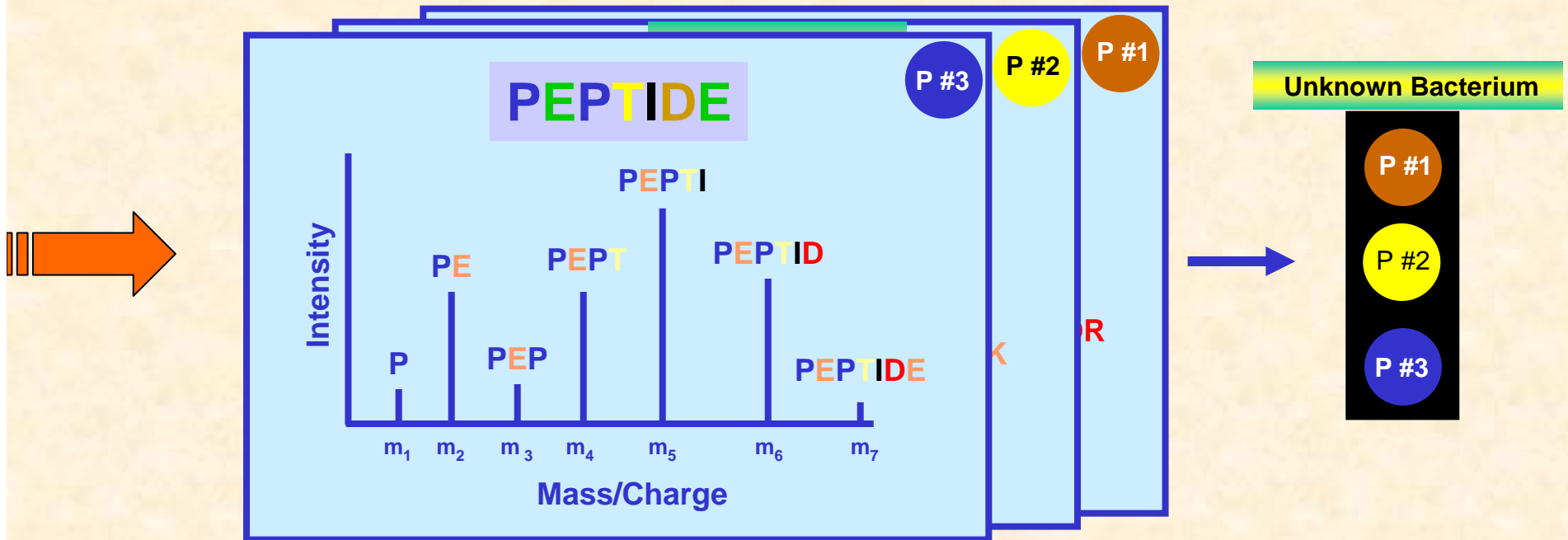


## Approach: (2) Tandem Mass Spectrometry of Peptide Ions

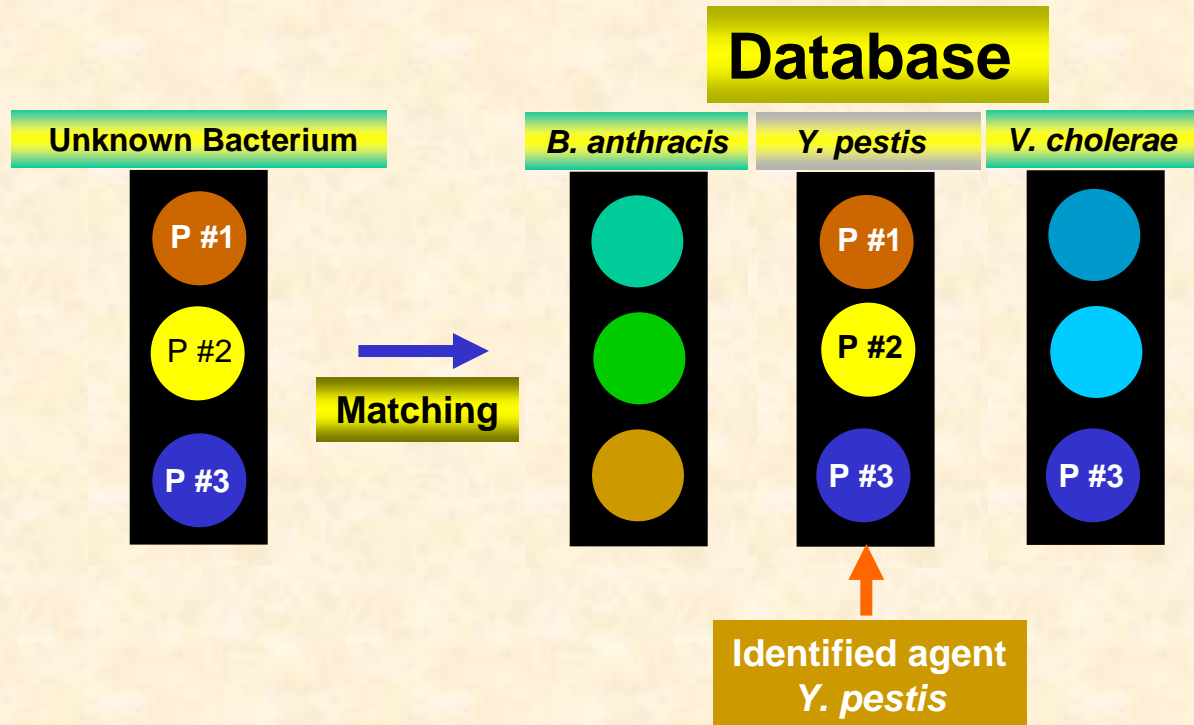


## Approach: (3) Sequencing of Peptide Ions

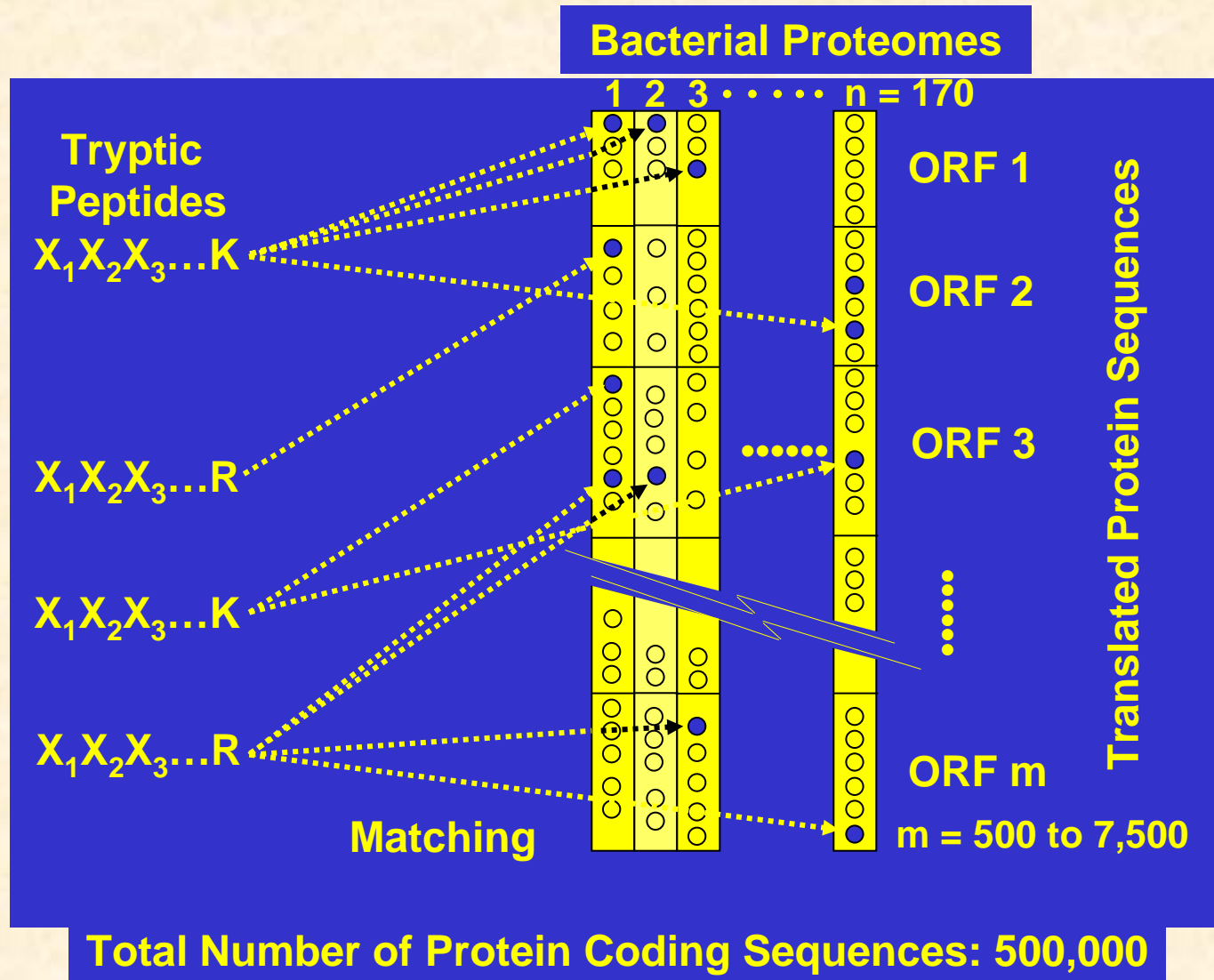
### MS/MS Spectra of Peptide Ions



## Approach: (4) Matching of Identified Tryptic Peptides to Theoretical Peptides of Database Bacterial Proteomes

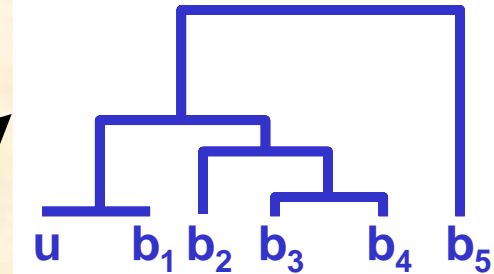
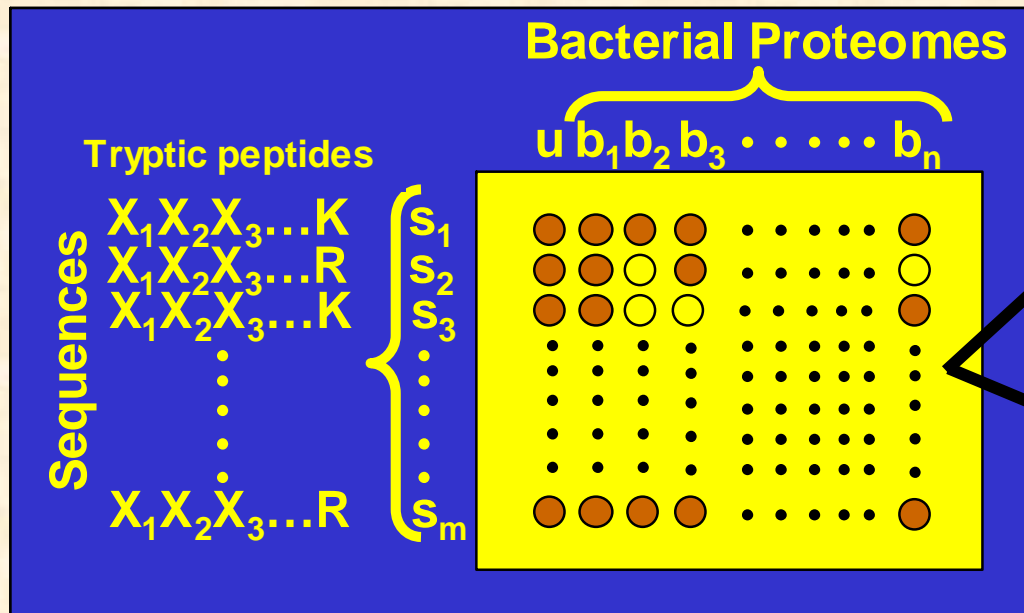


# Virtual Array of Peptide Sequences

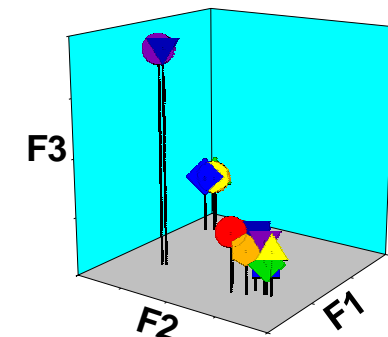


# Classification and Identification of Unknown Bacterium

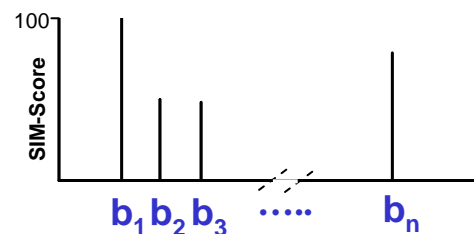
## Matrix of Assignments



## Cluster Analysis



## Principal Component Analysis



## Affinity Histogram

# Screen Capture Image Displaying Raw LC-MS/MS Data

SEQUEST Output Data

Discriminant Analysis Parameters

Filters

ABO Identifier

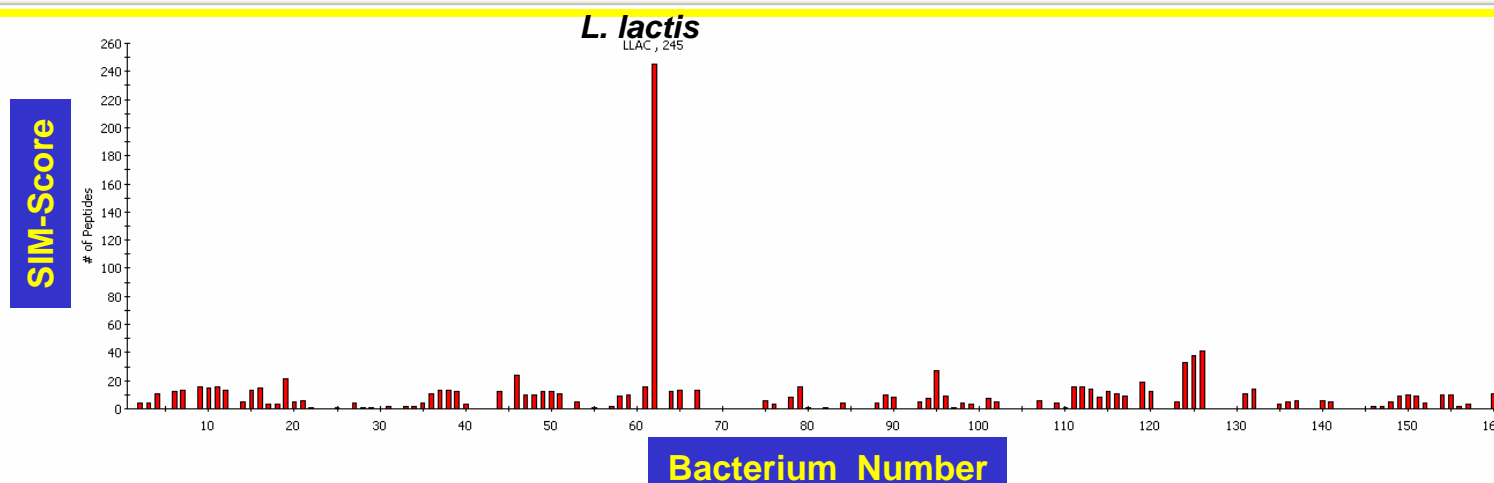
File Name

Sr.No	File Name	(M+H)	^M	^Cn	Xcorr	Sp	RSp	Reference	No	Peptide	No. Of AA	DF
1	FALL_A_Z_OUT_FOR_SAMIR\LL...	1352.7	0.2498	0.1160	1.453	459.0	000.000	XAXO_306	1	-LQALLAAAAVVR.-	13	0
2	FALL_A_Z_OUT_FOR_SAMIR\LL...	2030.3	1.4288	0.0000	0.000	048.5	000.693	BJAP_USDA110	1	-EHPEQWLWLRWR.-	14	0
3	FALL_A_Z_OUT_FOR_SAMIR\LL...	1954.2	0.3462	0.0630	0.675	102.9	000.000	LINT_56601	1	-EIQNWQIQYWNRF.-	14	0
4	FALL_A_Z_OUT_FOR_SAMIR\LL...	1689.9	1.255	0.1230	1.597	421.7	000.000	LLAC	1	-YGLTEMEVTEVFR.-	14	0
5	FALL_A_Z_OUT_FOR_SAMIR\LL...	1302.5	0.5575	0.0530	1.024	103.5	003.989	YPES_KIM	1	-LAALGATVHHR.-	12	0
6	FALL_A_Z_OUT_FOR_SAMIR\LL...	1953.1	1.0413	0.0330	0.697	038.3	002.303	LMON_EDGE	1	-DWHDTTEERDFWIR.-	14	0
7	FALL_A_Z_OUT_FOR_SAMIR\LL...	1899.1	0.2713	0.2540	0.738	040.8	002.303	SPNE_TIGR4	1	-NRVYDIDRYSYVK.-	14	0
8	FALL_A_Z_OUT_FOR_SAMIR\LL...	1439.6	0.5629	0.4080	3.653	870.6	000.000	LLAC	1	-VSPILGFTDQSK.-	13	1
9	FALL_A_Z_OUT_FOR_SAMIR\LL...	1844.1	0.3434	0.0490	1.118	060.7	003.401	AAEO_VF5	1	-LYYTHLSYLRNPF.-	14	0
10	FALL_A_Z_OUT_FOR_SAMIR\LL...	1846.0	0.7233	0.0330	1.503	060.7	003.296	VVUL_CMCP6 VVUL_Y3016	2	-FRDNFWDFGEFRVEK.-	14	0
11	FALL_A_Z_OUT_FOR_SAMIR\LL...	1957.3	0.0537	0.1890	0.838	015.8	001.609	WGLO	1	-VVMYINFYKFFHR.-	14	0
12	FALL_A_Z_OUT_FOR_SAMIR\LL...	1711.0	0.1139	0.0560	1.045	182.0	003.466	EFAE_V583	1	-FMTKEEAEQLVKEK.-	14	0
13	FALL_A_Z_OUT_FOR_SAMIR\LL...	1756.0	1.1425	0.1370	1.203	223.6	000.693	MGAL_R BANT_A2012 BA...	3	-YLQFSGQEKVYHK.-	14	0

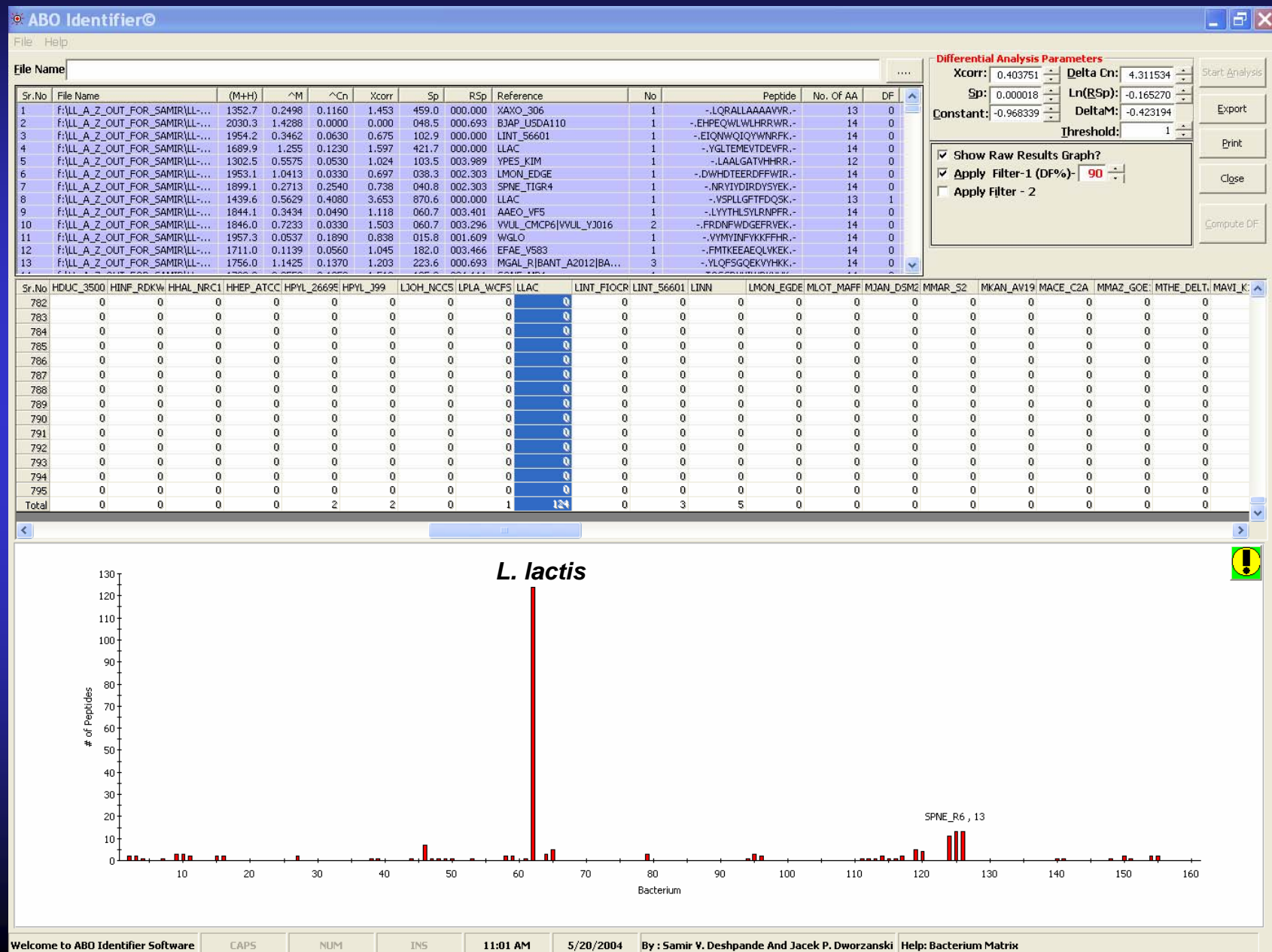
Sr.No	HDUC_3500	HINF_RDKW	HHAL_NRC1	HHEP_ATCC	HPYL_2669S	HPYL_999	LJOH_NCCS	LPLA_WCF5	LLAC	LINT_FIOCR	LINT_56601	LINN	LMON_EDGE	MLOT_MAFF	MJAN_DSM2	MMAR_S2	MKAN_AV19	MACE_C2A	MMAZ_GOE	MTHE_DELT	MAVI_KC
245	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
246	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
247	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
248	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
249	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
252	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
253	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
254	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
255	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
256	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
257	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
258	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
259	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
260	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Assignment Matrix

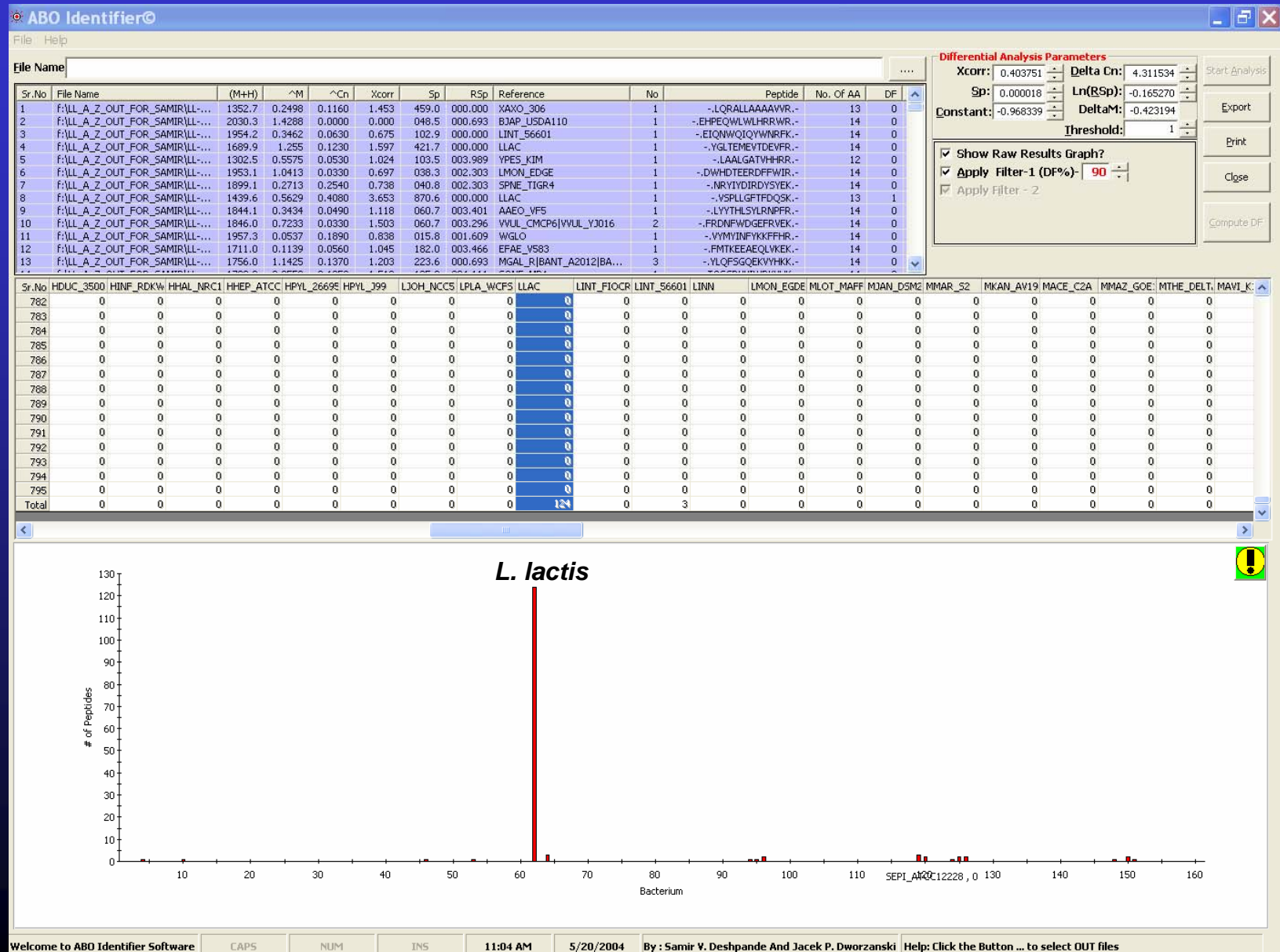




# Screen Capture Image Displaying Accepted Peptide Assignments (P = 90%)



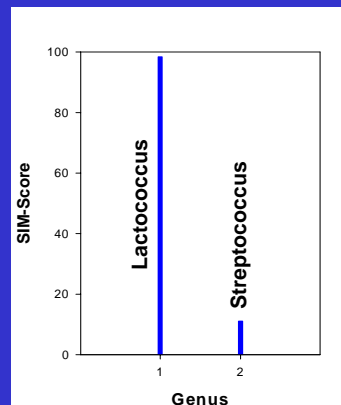
# Screen Capture Image Displaying Accepted Peptide Assignments After Removal of Degenerate Peptide Sequences



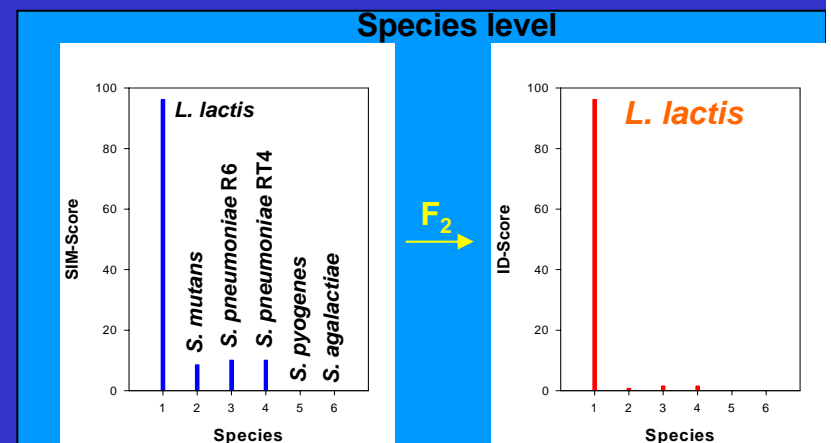
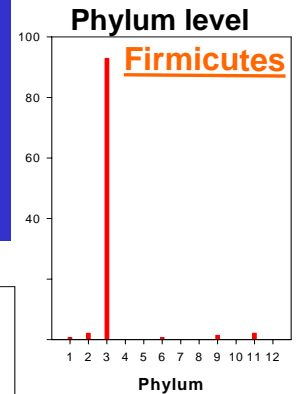
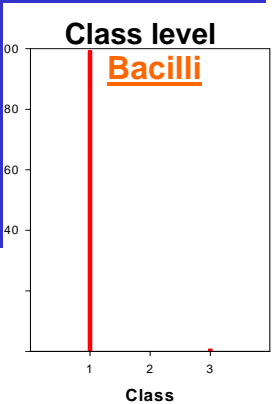
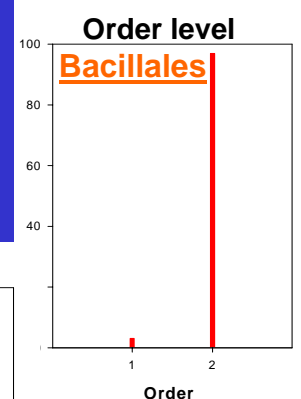
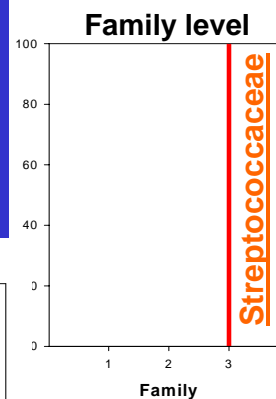
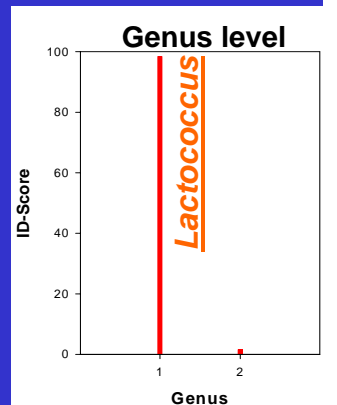


# Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides

*(Lactococcus lactis)*

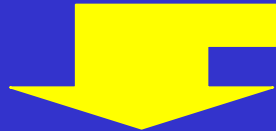


$F_2$

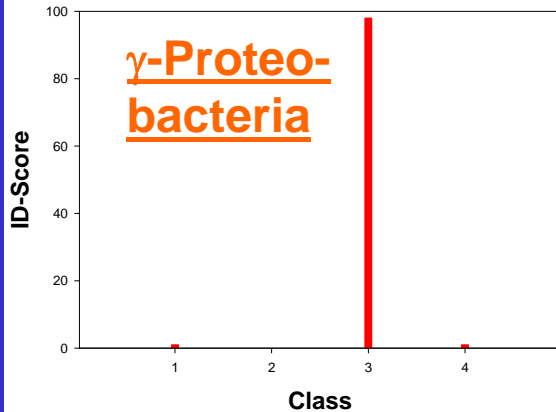


# Analysis of Bacterial Mixture

[*E. coli* (K-12) and *B. subtilis* (2:1), w:w]

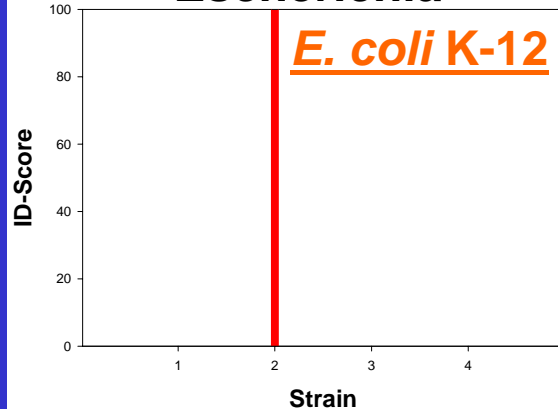


## Proteobacteria

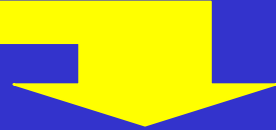
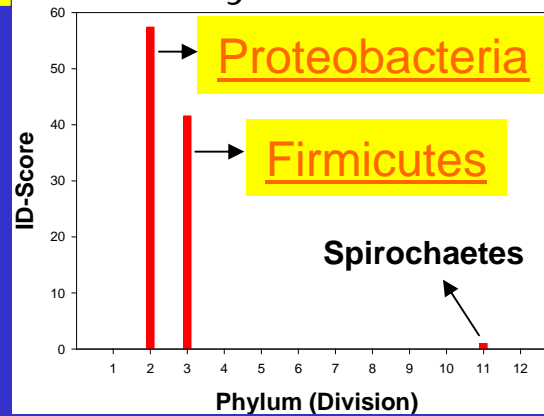


- ▼ Enterobacteriales
- ▼ Enterobacteriaceae
- ▼

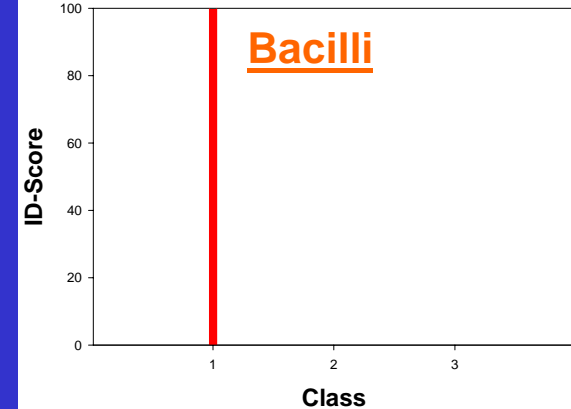
## Escherichia



## Phylum Level

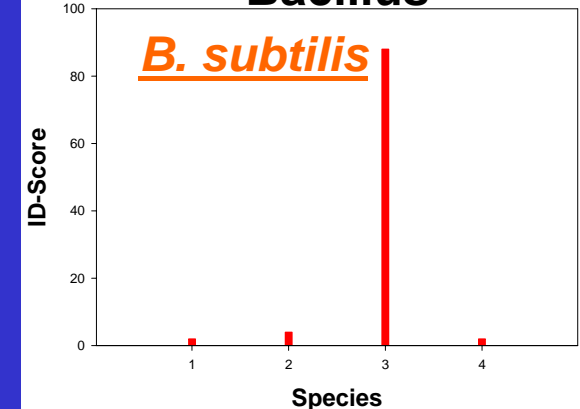


## Firmicutes



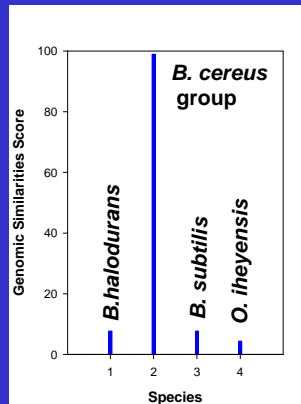
- ▼ Bacillales
- ▼ Bacillaceae
- ▼

## Bacillus

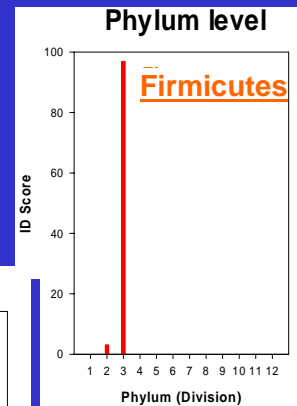
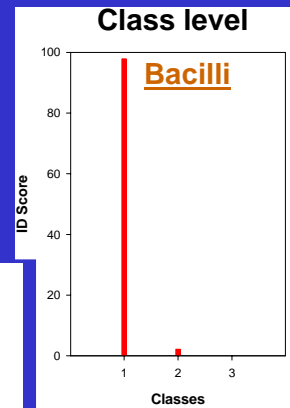
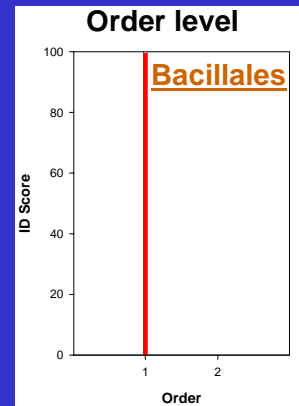
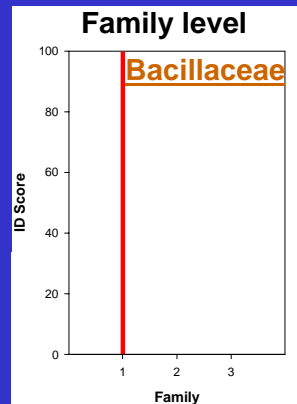
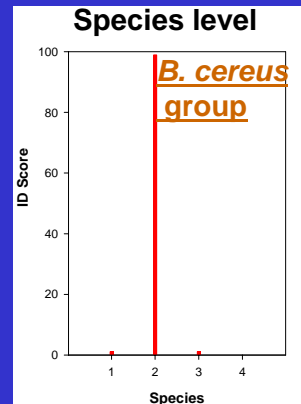


# Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides

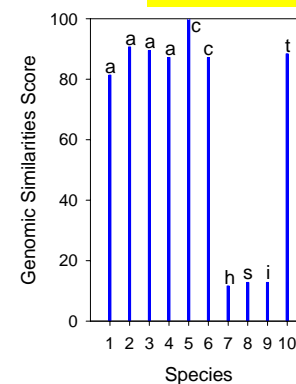
(*Bacillus cereus*)



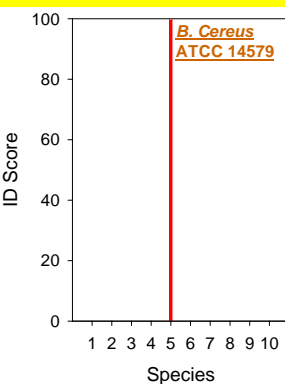
$F_2$



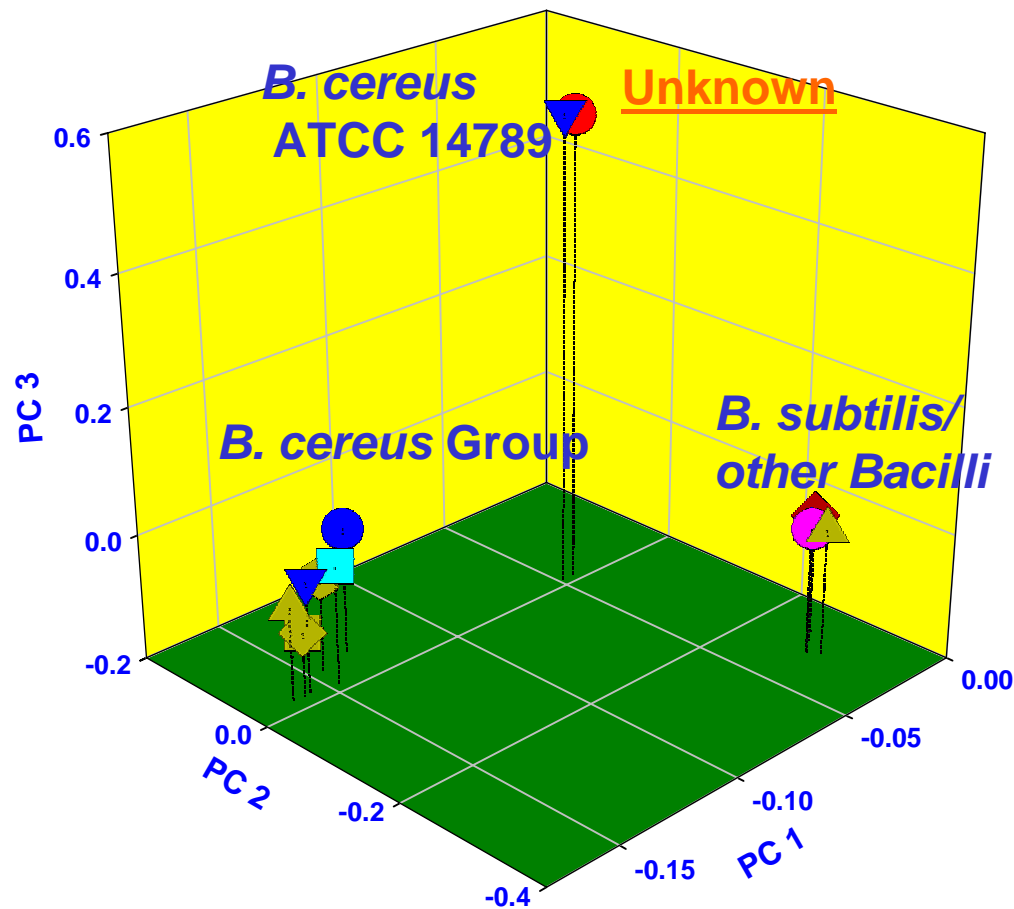
## Subspecies level (*B. cereus* group)



$F_2$

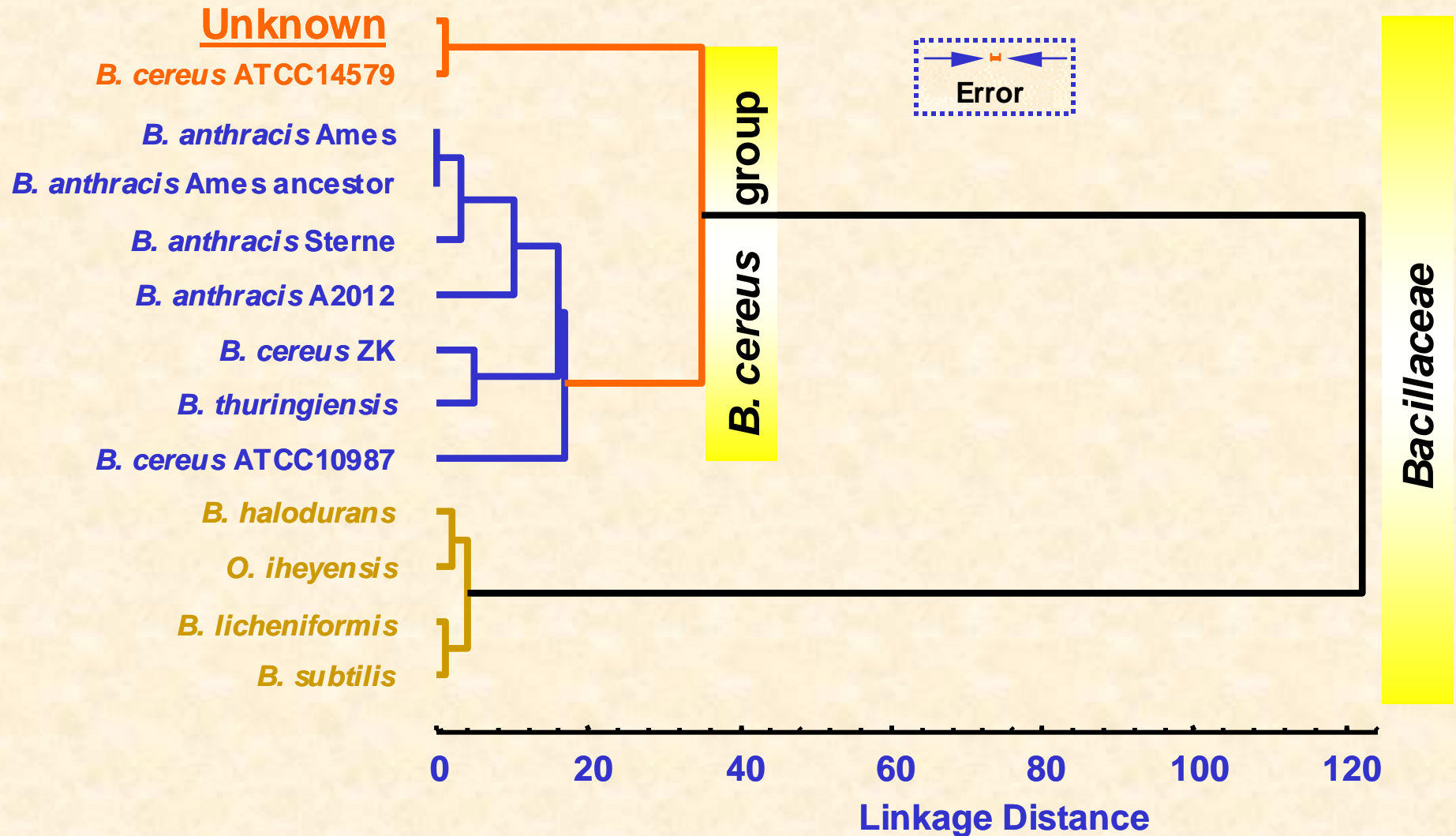


# Principal Component Analysis of Peptide Assignments



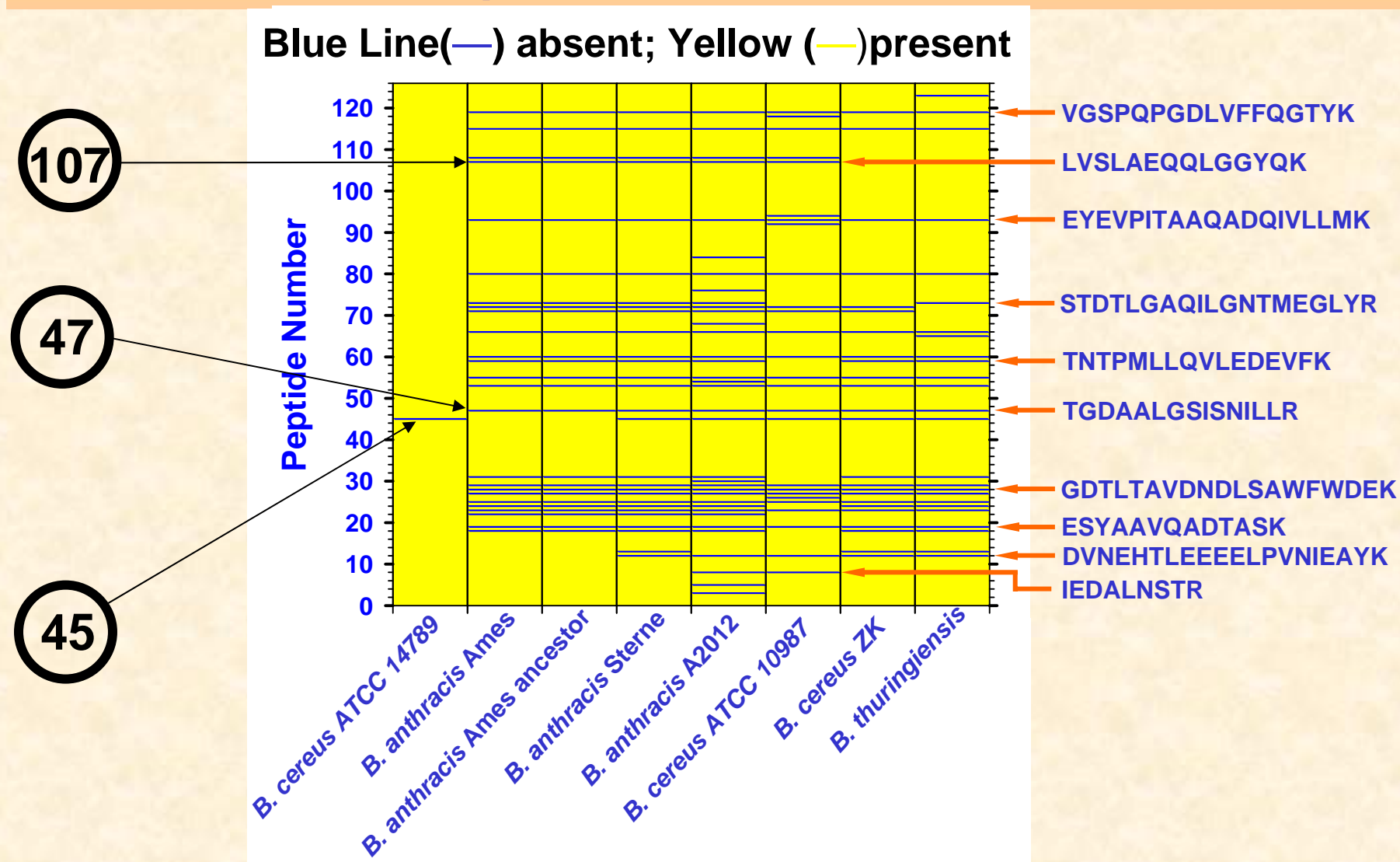
Representation of the database *Bacillaceae* species and unknown organism in the principal component space (PC 1, PC 2, PC 3) reflecting 80 % of the total information included in the assignment matrix of 125 amino acid peptide sequences to bacterial proteomes.

# Cluster Analysis of Peptide Assignments



Hierarchical clustering of *Bacillaceae* species in 125-dimensional space of peptide sequences.  
(Complete linkage; squared Euclidean distances)

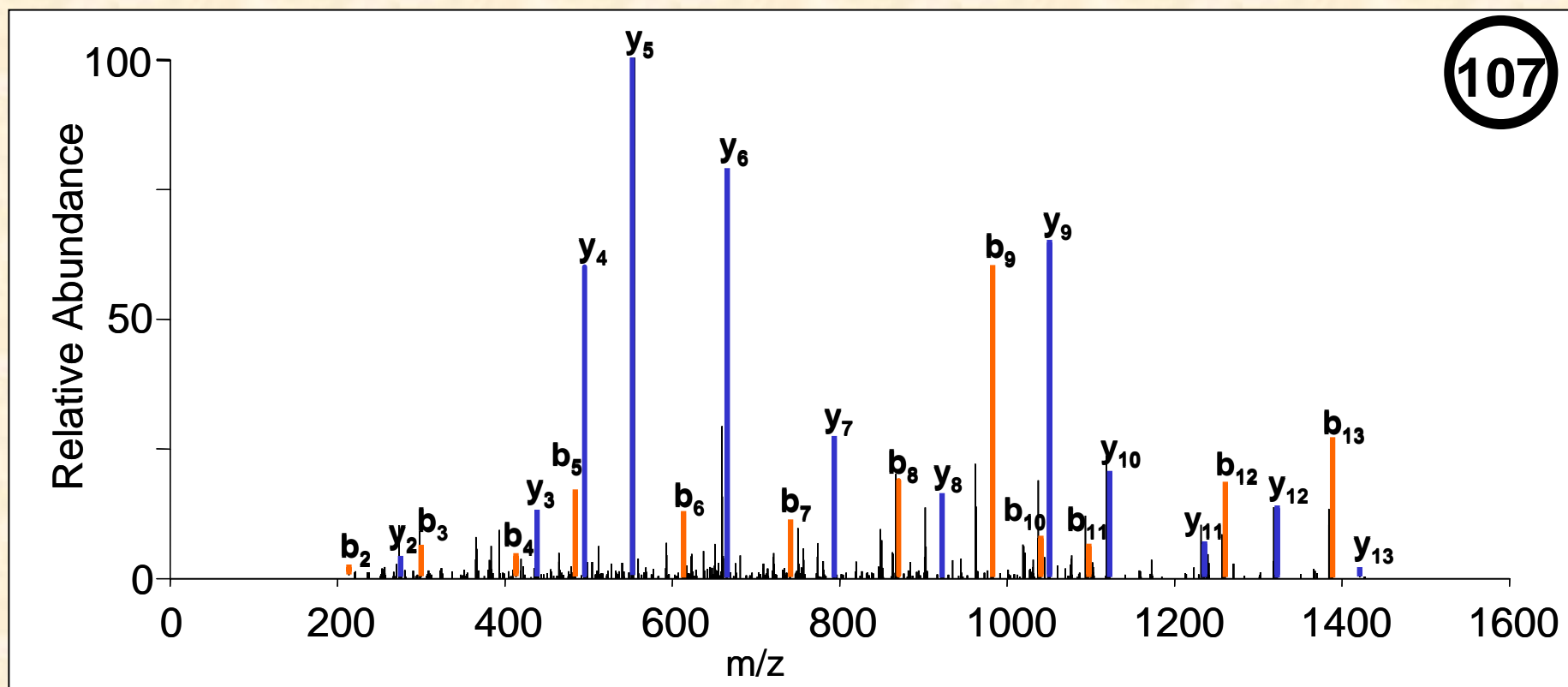
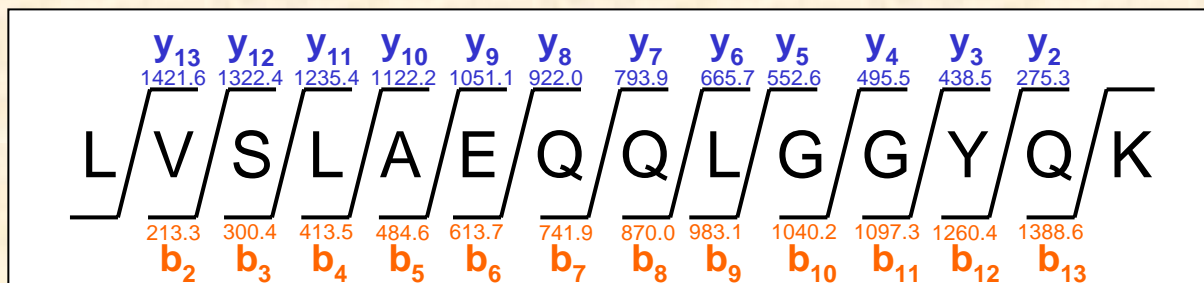
# Differences in Proteome Composition Between an Unknown Sample and Database Bacteria



Unknown sample: correctly identified as *B. cereus* ATCC 14789

# Product Ion Mass Spectrum of a Peptide Ion

Amino Acid Sequence Information Obtained in Less than 1 second



# Discriminative Power of DNA and Protein Sequences

Amino Acid  
Sequence From the  
MS/MS Spectrum

.....L V S L A E Q Q L G G Y Q K

*B.cereus*.ATTC 14579.....GAT CAA AGT GAT CGA CTC GTT GTT AAT CCG CCA ATG GTT TTT  
*B.anthraxis* A2012 .....GAT CAA AGT GAT CGA CTC GTT GTT AAT CCG CCA AAG GTT TTT

SASP-21402693

.....L V S L A E Q Q L G G F Q K

DNA Sequences

Amino Acid Sequences of  
Matching Peptides Found in  
the Database

Small, Acid-Soluble Spore Proteins

Accession		LVSLAEQQLGG[ ]YQK	
1430022721	52	.....	65 <a href="#">Bacillus cereus ATCC 14579</a>
<a href="#">21402693</a>	52	.....	65 <a href="#">Bacillus anthracis str. A2012</a>
<a href="#">42783829</a>	52	.....	65 <a href="#">Bacillus cereus ATCC 10987</a>
<a href="#">21401004</a>	56	.....	69 <a href="#">Bacillus anthracis str. A2012</a>
<a href="#">3688809</a>	54	..QM..M..	67 <a href="#">Bacillus firmus</a>
<a href="#">15613701</a>	53	..AM.....	67 <a href="#">Bacillus halodurans</a>
<a href="#">15615764</a>	54	..AM..M..	67 <a href="#">Bacillus halodurans</a>
<a href="#">134232</a>	56	..Q.....	66 <a href="#">Bacillus megaterium</a>
<a href="#">134239</a>	56	..Q.....	69 <a href="#">Bacillus megaterium</a>
<a href="#">134223</a>	56	..Q.....	69 <a href="#">Bacillus megaterium</a>
<a href="#">21399863</a>	54	..AM.....	64 <a href="#">Bacillus anthracis str. A2012</a>
<a href="#">30020123</a>	56	..AM.....	70 <a href="#">Bacillus cereus ATCC 14579</a>
<a href="#">30021203</a>	56	..AM.....	70 <a href="#">Bacillus cereus ATCC 14579</a>
<a href="#">42782199</a>	56	..AM.....	70 <a href="#">Bacillus cereus ATCC 10987</a>
<a href="#">21401007</a>	56	..AM.....	70 <a href="#">Bacillus anthracis str. A2012</a>
<a href="#">21398813</a>	54	..AM..S..	67 <a href="#">Bacillus anthracis str. A2012</a>
<a href="#">134230</a>	57	.....Q.....	70 <a href="#">Thermoactinomyces sacchari</a>



## Selected Peptide Sequences Discriminating Between an Unknown and Database Bacteria

Peptide #	Sequence	Protein
8	IEDALNSTR	60 kDa chaperonin GROEL
12	DVNEHTLEEEELPVNIEAYK	Hypothetical protein BC0479
19	ESYAAVQADTASK	Putative transcriptional regulator
29	GDTLTAVDNDLSAWFWDEK	Spore coat-associated protein N
45	KQPNFDDSSNFAK	Hypothetical protein BA 3347 [ <i>Bacillus anthracis</i> Ames]
47	TGDAALGSISNILLR	Flagellin
59	TNTPMLLQVLEDEVFK	Propionyl-CoA carboxylase biotin-containing subunit
73	STDTLGAQILGNTMEGLYR	Oligopeptide-binding protein oppA
93	EYEVPIAAQADQIVLLMK	IG hypothetical 17696
107	LVSLAEQQLGGYQK	Small acid-soluble spore protein
119	VGSPQPGDLVFFQGTYK	N-acetylmuramoyl-L-alanine amidase

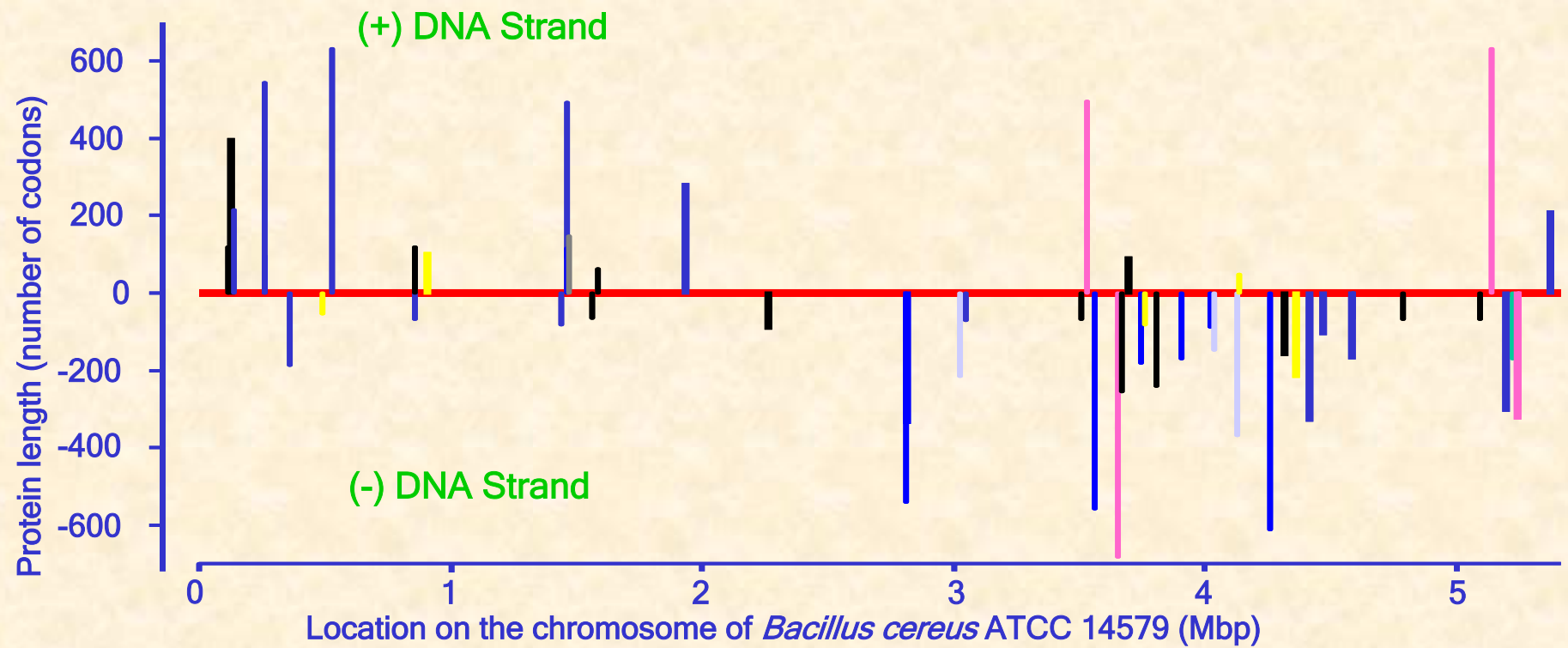
45



47



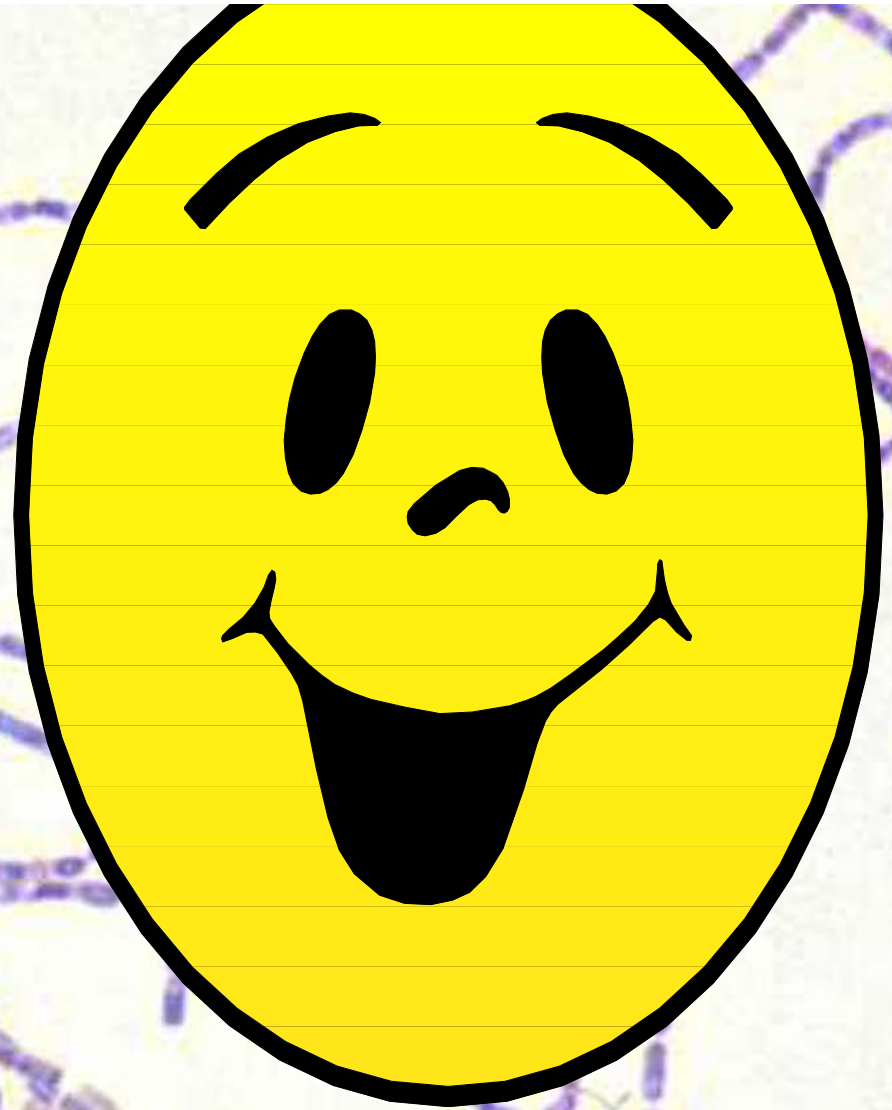
## Genes Encoding Identified Proteins



# CONCLUSIONS

The results demonstrate that mass spectrometry-based proteomics approach allows for :

- High confidence level classification and identification of bacteria based on genome traceable, proteomic similarities and differences between an analyzed microorganism and reference bacteria;
- Identification of pure cultures as well as mixtures of microorganisms.



Thank you !!!